DOCUMENT RESUME

ED 346 123 TM 018 376

AUTHOR Snyder, Patricia; Lawson, Stephen
TITLE Evaluating Statistical Significance Using Corrected and Uncorrected Magnitude of Effect Size Estimates.
PUB DATE Apr 92
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Data Interpretation; *Effect Size; *Estimation (Mathematics); *Measurement Techniques; Predictor Variables; Research Design; Research Methodology; Sample Size; Statistical Bias; *Statistical Significance
IDENTIFIERS *Association Strength; Linear Models; Preservice Teachers Dependent Variables

ABSTRACT

Magnitude of effect measures (MEMs), when adequately understood and correctly used, are important aids for researchers who do not want to rely solely on tests of statistical significance in substantive result interpretation. The MEM tells how much of the dependent variable can be controlled, predicted, or explained by the independent variables. Why methodologists encourage the use of MEMs as interpretive aids is described, and different types of measures are discussed. Correction formulas are presented to attenuate statistical bias in MEMs. MEMs are broadly grouped into measures of effect size and measures of associative strength. Several dozen computational indices are discussed in the literature concerning MEMs. The paper reviews the following categories: (1) biased versus unbiased magnitude of association measures; (2) population versus sample indices; (3) indices for fixed versus random-effect design models; (4) univariate versus multivariate magnitude of effect measures; and (5) equivalent measures from varying perspectives of the general linear model. Eight tables illustrate formula effects for differing sample and effect sizes. Several cautions against the indiscriminate use of these measures are offered. There is a 39-item list of references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED346123

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

PATRICIA SNYDER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Evaluating Statistical Significance Using Corrected and Uncorrected Magnitude of Effect Size Estimates

Patricia Snyder and Stephen Lawson

School of Allied Health Professions

Louisiana State University Medical Center

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, April 22, 1992.

TM018376

BEST COPY AVAILABLE

Abstract

Magnitude of effect measures, when adequately understood and correctly used, are important aids for researchers who do not want to place a sole reliance on tests of statistical significance in substantive result interpretation. We describe why methodologists encourage the use of ME measures as interpretation aids and discuss different types of ME measures. We discuss correction formulas developed to attenuate statistical bias in ME measures, and illustrate the affect of these formulas on different sample and effect sizes. Finally, we discuss several cautions against the indiscriminate use of these measures and offer reasons why we believe ME measures, like all substantive result interpretation aids, are useful only when their strengths and limitations are understood by researchers.

Evaluating Statistical Significance Using Corrected and Uncorrected Magnitude of Effect Size Estimates

Statistical significance testing is a prominent feature of data analytic traditions in the social sciences. For many years, methodologists have debated and discussed at length what statistical significance testing means and how it should be used in the interpretation of substantive results (e.g., Bakan, 1966; Carver, 1978; Chow, 1988; Greenwald, 1975; Hays, 1963; Huberty, 1987; Meehl, 1978; Morrison & Henkel, 1970; Thompson, 1989). The authors of a series of articles appearing in recent editions of the American Psychologist continue the discussion of statistical significance testing and common, persistent misconceptions associated with this tradition (e.g., Cohen, 1990; Kupfersmid, 1988; Rosnow & Rosenthal, 1989). These authors, as their predecessors, continue to educate us about the proper uses and interpretations of statistical significance tests and suggest several important alternatives to these tests.

The purpose of this paper is to discuss the use of magnitude of effect (ME) measures as one alternative for augmenting interpretation of findings beyond traditional tests of statistical significance. We describe why methodologists encourage the use of ME measures as interpretation aids and discuss different types of ME measures. We examine how these measures can contain statistical and methodological bias. We discuss correction formulas developed to attenuate statistical bias in ME measures, and illustrate the affect of these formulas on different sample and effect sizes.

Finally, we discuss several cautions against the indiscriminate use of these measures and offer reasons why we believe ME measures, like all substantive result interpretation aids, are useful only when their strengths and limitations are understood by researchers.

Why Use Magnitude of Effect Measures?

Methodologists suggest researchers should employ ME measures in result interpretation to highlight the distinction between "statistical" and "practical" significance (cf., Thompson, 1988). Results that are statistically significant are not always noteworthy in a practical sense. Carver (1978) and Rosnow and Rosenthal (1989) stated that a magnitude of effect measure tells us something very different from a p value. A magnitude of effect measure tells us how much of the dependent variable can be controlled, predicted, or explained by the independent variable(s).

Statistical Significance versus Importance

In many applied studies, researchers are interested in accounting for the largest proportion of variance or examining the amount of variance a variable explains of another variable (Chow, 1988; O'Grady, 1982). In these types of studies, use of ME measures can assist the researcher in clarifying whether statistically significant findings are of practical or meaningful significance within the context of an empirical investigation. For example, use of an instructional method that increases the performance of an experimental group on a dependent measure by 5 points over a control group results in statistically significant findings, if sample size is large enough. Whether or not such a 5

point difference (i.e., magnitude of effect measure) between the groups is meaningful from an instructional standpoint depends on many factors besides the statistically significant p value.

It is critical that researchers recognize that a small p value does not necessarily imply that the strength of the relation between the independent and dependent variables in a particular study is large. Rosnow and Rosenthal (1989) noted in the case of F ratios, a numerator mean square (MS) may be large relative to a denominator MS because the magnitude of the effect (i.e., sum of squares) is large, the number of conditions is large, or because both values are large.

Systematic examination of the magnitude of the effect also can assist the researcher in determining how much sample size is influencing results. Although achieving statistical significance is a function of at least seven interrelated study features (Schneider & Darcy, 1984), sample size is the primary influence on whether or not results will be statistically significant. Craig, Eison, and Metze (1976) stated, "Given a large enough sample size, a significant result may be identified when there is very little association between the independent and dependent variables" (p. 280). As Hays (1963) noted:

> The occurrence of a significant result says nothing at all about the strength of association between treatment and scores. A significant result leads to the inference that some association exists, but in no sense does this mean that an important degree of association necessarily

> exists. Conversely, evidence of a strong statistical association can occur in data even when the results are not significant. The game of inferring the true degree of statistical associations has a joker: this is the sample size. The time has come to define the notion of the strength of the statistical association more sharply, and to link this idea with that of the true difference between population means. (p. 324)

Sample Size and Statistical Significance Within the Context of "Fixed" Magnitudes of Effect

Maxwell and Delaney (1990) offered the formula

test statistic = size of effect x size of study

to illustrate the factors that multiply together to determine the value of the test statistic. The "size of study" term is usually some function of the sample size, usually a degrees of freedom term. Use of this formula to evaluate the magnitude of the test statistic highlights the point that a researcher can achieve statistical significance (i.e., larger test statistic) by increasing effect size or study size. Examining the size of a sample within the context of a "fixed" magnitude of effect allows the researcher to answer "what if" questions about statistical significance. That is, researchers can ask, "For a fixed magnitude of effect at what smaller sample size would my resulting test statistic no longer be significant?". Alternatively, researchers can ask, "For a fixed magnitude of effect at what larger sample size would my test statistic become significant?"

In Table 1 we illustrate this application. The table presents tests of statistical significance associated with varying sample sizes and a fixed magnitude of effect (33%). The results presented in Table 1 can be viewed as either a multiple regression analysis, involving two predictor variables, or an analysis of variance (ANOVA) involving an omnibus test of differences in three means in a one-way design involving a factor called "A". Both analyses invoke the F test statistic. The magnitude of effect measure in the multiple regression analysis would be called the coefficient of determination or squared multiple correlation coefficient ($R^2 = SOS_{explained}/SOS_{total}$). In the analysis of variance, this ME would be called eta squared or the correlation ration ($SOS_A/(SOS_A + SOS_{Error})$. For the 33% fixed effect size reported in Table 1, the result becomes statistically significant when there are somewhere between 13 and 23 subjects in the analysis.

Insert Table 1 about here

By systematically examining the influence that adding or removing subjects has on statistical significance, researchers can develop an increased understanding of how sample size is affecting statistical significance. Substantive results can then be interpreted accordingly. For example, a result may not be statistically significant with 13 subjects in the analysis, although the ME is 30%. By adding two subjects to the study, the results become statistically significant with 15 instead of 13

subjects in the analysis. This does not mean that the results without these two subjects are not clinically noteworthy. Result interpretation should always include an explicit analysis within the context of the ME of when statistically nonsignificant results would become statistically significant, or when statistically significant results would no longer be statistically significant (Thompson, 1988).

We acknowledge that the purpose of interpreting statistical significance within the context of a "what if" fixed magnitude of effect approach is not to identify the **exact** results that would occur with a different size sample. In applied settings, the addition or removal of subjects in an analysis would likely alter the explained and unexplained variance and therefore the obtained magnitude of effect. However, we maintain that this "what if" approach can provide researchers with important insight regarding the effects of sample size on statistically significant, or nonsignificant, results that may not be entirely evident from p values alone.

Types of Magnitude of Effect Measures

Notwithstanding the importance of magnitude of effect measures as important result interpretation aides, several details about these measures that could confuse the "bright but naive" researcher immediately become obvious as one begins to explore the magnitude of effect literature. In our view, these details merit some discussion. We believe an understanding of these details leads the researcher to employ magnitude of effect measures in a more

informed manner.

Categories of Magnitude of Effect Measures

The first detail is to clarify the terms used to characterize these measures. Murray and Dosser (1987) and O'Grady (1982) noted authors may use any of the following terms to refer to these measures: (a) measures of the magnitude of the effect, (b) measures of the magnitude of the experimental effect, (c) measures of explained variance, (d) effect size measures, (e) measures of the strength of relation, or (f) measures of the strength of association.

Maxwell and Delaney (1990) attempted to rectify some of this semantic confusion by organizing magnitude of effect measures under two broad categories, measures of effect size and measures of association strength. These authors included mean difference indices, estimated effect parameter indices, and standardized differences between means (e.g., measures used in meta-analysis such as Cohen's *d*, Cohen, 1988) in the measure of effect size category. Thus, the first category includes those measures that involve examining differences between means.

The second category of magnitude of effect measures includes indices that involve proportions of variance, how much of the variability in the dependent variable(s) is associated with the variation in the independent-variable(s). The magnitude of measures of association range from 0 to 1. Examples include indices such as eta squared (Fisher, 1932), partial eta squared (Cohen, 1973), omega squared (Hays, 1963), epsilon squared (Kelley,

1935), R^2 (Stevens, 1986), partial R^2 (Cohen & Cohen, 1983), the Wherry formula (Wherry, 1931), the Herzberg formula (Herzberg, 1969), and the Lord formula (Lord, 1950). Authors present different computational formulas in the literature for these various indices.

Computing Magnitudes of Effect

The different computational formulas constitute a **second** detail that researchers need to understand when employing magnitude of effect measures to assist in substantive result interpretation. There are several dozen computational indices discussed in the magnitude of effect literature (see Carter, 1979; Camp & Maxwell, 1983; Maxwell, Camp, & Avery, 1981; Maxwell & Delaney, 1990; O'Grady, 1982 for reviews). In the sections that follow, we will compare and contrast several magnitude of association strength indices. Readers interested in more detail regarding measures of effect size can consult Cohen (1988), Glass, McGaw, and Smith (1981), and Maxwell and Delaney (1990).

Measures of association strength discussed in the magnitude of effect literature may vary along at least four dimensions: (a) "biased" or "unbiased" indices, (b) "population" or "sample" indices, (c) "fixed-" or "random-effect" design model indices, and (d) "univariate" or "multivariate" indices. Finally, although it is widely recognized that all univariate parametric techniques are special cases of regression (Cohen & Cohen, 1983; Thompson, 1988) various indices of measures of association for ANOVA and regression, although similar from a computational and conceptual

perspective, rarely are presented as analogues in a general linear model sense (notable exceptions are reviews by Maxwell & Delaney, 1990; and O'Grady, 1982). The next several sections include relatively explicit descriptions of different measures of association strength in terms of their derivations, underlying assumptions, and applications.

Biased versus unbiased magnitude of association measures. In its simplest form, a measure of association can be expressed as the ratio of explained variance to total variance. Eta squared (ANOVA) and R^2 (regression) discussed previously and illustrated in Table 1 are examples of these simple measures of association strength. These indices are positively biased. This means that these measures have a tendency to overestimate systematically the proportion of variability that might be explained in the population or in future samples. These estimates are based directly on the data obtained from a particular study sample.

As Stevens (1986) explained, the overestimates in these simple measures of association strength result from the mathematical maximization principle operating in all general linear model analyses. The linear combination of x's that is maximally correlated with some y is sought when sample results are analyzed. Minimizing the sum of squared errors is equivalent to maximizing the correlation between x and y scores. Any variation in the study samples due to sampling error will cause the measure of association to be positively biased. Even if there is no systematic relation between x and y in the population ($R = 0$), the sample R^2 or eta

squared is unlikely to ever be exactly zero.

O'Grady (1982) noted that the amount of bias in a measure of association strength varies according to a number of factors including the: (a) reliability of the measurement instruments, (b) research questions posed, (c) sample size, (d) number of predictor or independent variables under investigation in a particular study, (e) heterogeneity of the study sample, and (f) type of design used to investigate a particular research question. In general, experimental studies and more homogeneous samples result in smaller measures of association. Studies that involve small subject to variable ratios (5:1 or less) will usually contain a more noteworthy amount of positive bias. The reliability of the measurement instrument sets an upper bound on the magnitude of the effect that can be realized in any substantive investigation. O'Grady illustrates how even the types of research questions posed can affect the magnitude of the measure of association.

Methodologists developed what are termed "unbiased" measures of association strength estimates in an effort to **statistically** correct for the bias operating in the simpler indices. These "unbiased" estimates allow the researcher to make **inferences** regarding expectations for the effect size that might be realized in the population or in future samples. Unbiased measures of association strength estimates that we will compare and contrast in this paper include: (a) omega squared, (b) epsilon squared, (c) the Wherry formula, (d) the Herzberg formula, and (e) the Lord formula. Table 2 contains the formulas researchers use to compute these

estimates.

Insert Table 2 about here

Population versus sample indices. Formulas designed to develop estimates of measures of association strength likely to be realized in other samples result in more conservative estimates than those for populations. The Herzberg and Lord formulas are designed to correct for bias in developing estimates likely to be realized in future samples. The measures of association strength values derived using these two formulas will always be more conservative (i.e., smaller) than omega squared, epsilon squared, or the Wherry formula. These latter three formulas are designed to estimate the measure of association strength likely to be realized in the population.

The issue of whether to use a bias-correction formula designed to estimate measure of association strength likely to be realized either in other samples or in the population depends on the types of generalizations the researcher wishes to make. Herzberg (1969) offers the following caveat, "In applications, the population regression function [i.e., effect size estimate] can never be known and one is more interested in how effective the sample regression function is other samples" (p. 4). Most researchers ground their work in empirical findings from previous samples and usually desire that their work generalize to future samples. There may be occasions, however, when the researcher desires an estimate

designed for developing population expectations, use of the population bias-correction formulas would be appropriate in these instances.

Indices for fixed- versus random-effect design models. Different formulas are used to estimate measures of association strength likely to be realized in future samples or in the population for fixed- versus random-effect design models. Fixed models assume that levels in factors are fixed in an ANOVA design or the values of the predictors are fixed in a regression model, they **do not** change from one replication of the study to another. Levels of factors in random-effect design models are selected as a result of a random process in ANOVA or the predictors are random variables in the correlational model, they **can** change from one replication of the study to another.

The bias correction formulas for random-effect design models differ from fixed-effect formulas. For example, Howell (1982) presented omega squared formulas for fixed- and random-effect design models in ANOVA. Stevens (1986) presented different formulas for the Herzberg correction for fixed- and random-effect design models. Both of these authors illustrate that using a fixed-effect formula to develop estimates of magnitude of association strength results in estimates that differ from those developed using random-effect formulas. Table 2 presents examples of the fixed- and random-effect formulas for omega squared and the Herzberg formula.

Kern and Lewis (1979), Murray and Dosser (1987), and O'Grady

(1982) emphasized the importance of researchers recognizing the assumed underlying design model when selecting a bias-correction formula. The way in which the underlying design model is constructed has implications for the magnitude of the measure of association strength we derive from our data, and the generalizations we might draw from our study (Howell, 1982).

Univariate versus multivariate magnitude of effect measures. In this paper we are discussing univariate ME measures. Formulas are available, however, to evaluate the ME in the multivariate case (Huberty, 1972). As Stevens (1986) noted, "The multivariate measure D^2 is a natural squared generalization of the univariate measure d, where the means have been replaced by the mean vectors and s (standard deviation) has been replaced by its squared multivariate generalization of within variability, the sample covariance matrix **S**" (p. 140). Pedhazur (1982) demonstrates how 1 - lambda is the multivariate analogue to R^2 or eta squared. Obviously, researchers asking multivariate questions will need to employ magnitude of effect measures that are consistent with their multivariate view of the research problem.

Equivalent measures from varying perspectives of the general linear model. Measures of strength of association developed from one perspective of the general linear model will have counterparts in the other analytic techniques that spring from the same source (Maxwell et al., 1981). Epsilon squared for fixed-effect design models, developed from the ANOVA perspective, yields numerically equivalent values to the Wherry formula developed from the

regression perspective. Omega squared for the random model (ANOVA) and the squared intraclass correlation (correlational model) also yield numerically equivalent values. Maxwell et al. (1981) noted that researchers should view all measures of association strength from the perspective of the general linear model. In this way, they will avoid maintaining an artificial and unwarranted distinction between the various measures.

Answering "What If" Questions Using Corrected and Uncorrected Measures of Association Strength Indices

How much difference does it make which measure of association strength index a researcher chooses to use to assist in evaluating the magnitude of effect? The answer is, it depends. When sample sizes and effect sizes are large, the biased as well as the unbiased correction formulas produce similar estimates. Thompson and Melancon (1990) noted "Thus, with a very large effect size, or a large sample size, or both, it will matter less which, if any, statistical corrections the researcher applies in estimating effect sizes" (p. 11). However, when effect sizes and sample sizes become smaller the statistical corrections tend to be larger (Thompson, 1990). How large the statistical corrections are varies according to the formula used to develop the estimate.

For heuristic purposes, we developed three source tables presented in Tables 3 - 5 to assist us in illustrating the amount of statistical correction across the various indices of measures of association strength in relation to size of the sample and magnitude of effect. These source tables, like Table 1 presented

previously, illustrate tests of statistical significance associated with varying sample sizes (10, 30, 50 and 100) and biased magnitudes of effect (i.e., eta squared or R^2). Table 3 contains a "fixed" biased magnitude of effect of .60, Table 4 contains a "fixed" magnitude of effect of .30, and Table 5 contains a "fixed" magnitude of effect of .10. The results presented in these tables can be viewed either as a multiple regression analysis involving two predictor variables, or an ANOVA involving an omnibus test of differences in three means in a one-way fixed effect model design involving a factor called "A". Both analyses invoke the F test statistic.

Insert Tables 3 - 5 about here

Tables 6 - 8 present the measures of association strength values for different indices and sample sizes based on the values contained in the three source tables using the formulas presented in Table 2. Eta squared or R^2 (i.e., the "fixed" effect across all sample sizes) values are presented first in these tables, followed by omega squared, epsilon squared, and values derived from the Wherry formula, the Herzberg formula, and the Lord formula.

Note that in Table 6 when the uncorrected ("fixed") measure of association strength is large (.60) and the sample size is relatively large (50 or 100), the bias correction formulas produce measures of association strength that closely approximate the uncorrected estimate (range .5488 to .5918). When sample sizes

decrease (10 or 30), the bias correction formulas produce more conservative estimates (range .2572 to .5704). Consistent with the observation made by Uhl and Eisenberg (1970), the Lord formula produces the most conservative estimates of measures of association strength across all sample sizes. This finding is not surprising given that both the Herzberg and Lord formulas are designed to develop estimates likely to be obtained in future samples, these corrections will always be more conservative than population estimates. The reader also will observe that the values in Table 6 for epsilon squared and the Wherry correction formula are identical. These two different formulas produce numerically equivalent values due to their common origin in the general linear model.

In Table 7, negative estimates are produced for the smallest sample size (10) when samples estimates (i.e., Herzberg and Lord) are calculated. A negative variance estimate value is mathematically impossible to obtain. Researchers report measures of association strength as zero in such instances (Cohen, 1973). Corrections are more conservative across all sample sizes when the biased effect size is moderate (.30) compared to large (.60).

The values of the estimates presented in Table 8 show how corrections produce much more conservative estimates when the biased measure of association strength is small (.10). At the smallest sample size (10) negative estimates are produced across all indices that correct for bias. Even at larger sample sizes (100) "shrinkage" of 44% to 80% occurs in the unbiased estimates

(range .0443 to .0814).

We note that the values presented in Tables 6 - 8 illustrate that using the omega squared formula results in more conservative bias correction estimates when compared to the epsilon squared formula or the Wherry formula. Thus, the omega squared formula develops more conservative population estimates and the Lord formula develops more conservative sample estimates.

Cautions Concerning Magnitude of Effect Indices

Most methodologists agree that ME indices can provide researchers with important additional insight into the meaning of substantive results, but only when used and interpreted properly (cf. Maxwell et al., 1981; Murray & Dosser, 1982; O'Grady, 1982). We review several cautions about these measures offered in the ME literature. An appreciation of these cautions could assist researchers in using these measures productively.

First, generalizations regarding a ME estimate are limited to those instances that involve the same levels of the independent variable and similar numbers and types of subjects. Statistical generalization is impossible for levels not included in a fixed-effects design (Maxwell et al., 1981, p. 531). In fixed-effect design models, it is not sufficient to state that the percentage of variance accounted for in the dependent variable by the independent variable is 30%. It is more accurate to state that the k **particular** levels of the independent variable account for 30% of variance in the dependent variable when n subjects of p type are assigned to each cell.

Second, experimenters can choose levels of treatment that are known to vary widely and increase the probability of obtaining a large value for the measure of association strength. Maxwell et al. (1981) noted, however, that this same problem also exists for hypothesis testing in fixed-effect design models. Once again, the researcher should explicitly acknowledge the levels included in a particular study and limit generalizations regarding the measure of association to these levels.

These first two cautions indicate that magnitude of effect measures are very context-dependent. As Rosnow and Rosenthal (1989) noted:

> There is a growing awareness in psychology that just about everything under the sun is context dependent in one way or another....Strength of effect measures are no exception, and it is therefore important to recognize how the study characteristics might influence the size of one's interpretation of the magnitude-of-effect estimate. (p. 1280)

Third, measures of association have been criticized as being only point estimates of population measures of association that ignore the sampling error of the resulting estimate (Fowler, 1985). Fowler discusses in detail why he believes that confidence intervals can and should be constructed for measures of association.

Fourth, because ME measures vary due to both statistical and methodological factors the issue of "large effects" versus "small

effects" becomes confounded. Researchers frequently debate the issue of "how large" a ME measure should be to be considered clinically noteworthy. Although Cohen (1988) provided some guidelines regarding "large", "medium", and "small" effects, the judgement regarding the clinical significance of a ME ultimately rests with the researchers's personal value system, the research questions posed, societal concerns, and the design of a particular study. Throughout this paper we emphasized that a number of factors influence the magnitude of measures of association strength. Setting arbitrary guidelines against which to measure the size of **particular** ME discounts the context-dependency of the investigative process.

Finally, methodologists caution that measures of association strength expressed as percentage of variance accounted for may not provide us with enough information about the differences in means brought about by the treatments. Rosnow and Rosenthal (1989) propose using indices such as binomial effect size displays (BESD) to further clarify the "meaningfulness" of obtained results. Recent comments in the American Psychologist (October 1991) discuss the issues and controversies associated with BESD in more detail.

Summary

Magnitude of effect measures, when adequately understood and correctly used, are important aids for researchers who do not want to place a sole reliance on tests of statistical significance in substantive result interpretation. We reviewed several types of measures of association strength indices in this paper and

discussed in what contexts these measures might best be utilized. We illustrated how different statistical correction formulas are more or less conservative under certain conditions.

We emphasized in the paper that ME measures are context specific and, as such, are influenced by a variety of factors within particular research situations. This implies that ME measures, just like tests of statistical significance, have certain limitations and biases. Researchers who recognize the usefulness and limitations of these and other interpretation aids are more likely to make informed use of them in the interpretation of substantive results. These aids are merely tools to assist the researcher in gaining a more informed analysis of **data**. The ultimate responsibility for developing a comprehensive analysis of the meaning of **results** rests with the researcher. As Murray and Dosser (1987) stated:

> In conclusion, there is no statistical substitute for the knowledge and experience of the researcher. This knowledge and experience should form the basis on which the researcher decides what size differences are important and what effects are practical in his or her particular area of research. (p. 72)

References

Bakan, D. (1966). The test of significance in psychological research. Psychological Bulletin, 66, 423-437.

Camp, C.J., & Maxwell, S.E. (1983). A comparison of various strengths of association measures commonly used in gerontological research. Journal of Gerontology, 38, 3-7.

Carter, D.S. (1979). Comparison of different shrinkage formulas in estimating population multiple correlation coefficients. Educational and Psychological Measurement, 39, 261-266.

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Chow, S.L. (1988). Significance test or effect size? Psychological Bulletin, 103, 105-110.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. Educational and Psychological Measurement, 33, 107-112.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New York: Lawrence Erlbaum.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Craig, J.R., Eison, C.L., & Metze, L.P. (1976). Significance tests and their interpretation: An example utilizing published research and omega squared. Bulletin of the Psychonomic

Society, 7, 280-282.

Fisher, R.A. (1932). Statistical methods for research workers. Edinburgh: Oliver and Boyd.

Fowler, R.J. (1985). Point estimates and confidence intervals in measures of association. Psychological Bulletin, 98, 160-165.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Greenwald, A. (1975). Consequences of prejudices against the null hypothesis. Psychological Bulletin, 82, 1-20.

Hays, W.L. (1963). Statistics for psychologists. New York: Holt, Rinehart, and Winston.

Herzberg, P.A. (1969). The parameters of cross-validation. Psychometrika Monograph Supplement, 16, 1-67.

Howell, D.C. (1982). Statistical methods for psychology. Boston: Duxbury.

Huberty, C.J. (1972). Multivariate indices of strength of association. Multivariate Behavioral Research, 7, 523-526.

Huberty, C.J. (1987). On statistical testing. Educational Researcher, 16, 4-9.

Kelley, T.L. (1935). An unbiased correlation ratio measure. Proceedings of the National Academy of Sciences, 21, 554-559.

Kern, G., & Lewis, C. (1979). Partial omega squared for ANOVA designs. Educational and Psychological Measurement, 39, 119-128.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Lord, F.M. (1950). Efficiency of prediction when a regression equation from one sample is used in a new sample. Educational Testing Service Research Bulletin, 50-110.

Maxwell, S.E., Camp, C.J., & Avery, R.D. (1981). Measures of strength of association: A comparative examination. Journal of Applied Psychology, 66, 525-534.

Maxwell, S.E., & Delaney, H.D. (1990). Designing experiments and analyzing data: A model comparison perspective. Belmont, CA: Wadsworth.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Morrison, D.E., & Henkel, R.E. (Eds.). (1970). The significance test controversy. Chicago: Aldine.

Murray, L.W., & Dosser, D.A. (1987). How significant is a significant difference? Problems with the measurement of magnitude of effect. Journal of Counseling Psychology, 34, 68-72.

O'Grady, K.E. (1982). Measures of explained variance: Cautions and limitations. Psychological Bulletin, 92, 766-777.

Pedhazur, E.J. (1982). Multiple regression in behavioral research: Explanation and prediction. Fort Worth, TX: Holt, Rinehart and Winston.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Schneider, A.L., & Darcy, R.E. (1984). Policy implications of using significance test in evaluation research. Evaluation Review, 8, 573-582.

Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thompson, B. (1988). Common methodology mistakes in dissertations: Improving dissertation quality. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 301 595).

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement in Counseling and Development, 22, 2-5.

Thompson, B. (1990). Finding a correction for the sampling error in multivariate measures of relationship: A Monte Carlo study. Educational and Psychological Measurement, 50, 15-31.

Thompson, B., & Melancon, J.G. (1990, November). Bootstrap versus statistical effect size correction: A comparison with data from the Finding Embedded Figures Test. Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans.

Uhl, N., & Eisenberg, T. (1970). Predicting shrinkage in the multiple correlation coefficient. Educational and Psychological Measurement, 30, 487-489.

Wherry, R.J. (1931). A new formula for predicting the shrinkage of

the coefficient of multiple correlation. Annals of Mathematical Statistics, 2, 440-451.

Table 1

Statistical Significance at Various Sample Sizes for a Fixed Effect Size (33%)

Source	SOS	r^2	df	MS	Fcalc	Fcrit	Decision
			n = 4				
Explained	**337.2**	**0.336**	**2**	**168.6**	0.253	200.00	NR
Unexplain	**665.1**		1	665.1			
Total	**1002.3**		3	334.1			
			n = 13				
Explained	**337.2**	**0.336**	**2**	**168.6**	2.535	4.10	NR
Unexplain	**665.1**		10	66.51			
Total	**1002.3**		12	83.525			
			n = 23				
Explained	**337.2**	**0.336**	**2**	**168.6**	5.070	3.49	REJ
Unexplain	**665.1**		20	33.25			
Total	**1002.3**		22	45.559			
			n = 33				
Explained	**337.2**	**0.336**	**2**	**168.6**	7.605	3.32	REJ
Unexplain	**665.1**		30	22.17			
Total	**1002.3**		32	31.322			

Note. Entries in **bold** remain fixed for the purposes of these analyses. From Thompson (1989), with permission.

Table 2

Formulas for Computing Measures of Association Strength

Index	Formula
Eta squared (R^2)	$^{a}SOS_{explained} / SOS_{total}$
Omega squared	$SOS_{explained} - (([^{b}v-1) * {}^{c}MS_{error}) / SOS_{total} + MS_{error}$
Omega squared*	$(MS_{explained} - MS_{error}) / MS_{explained} + (([^{d}n-1) * MS_{error})$
Epsilon squared	$SOS_{explained} - ((v-1) * MS_{error} / SOS_{total}$
Wherry	$1 - ((n-1) / (n-{}^{e}k-1)) * (1 - R^2)$
Herzberg	$1 - ((n-1) / (n-k-1)) * ((n+k+1) / n) * (1 - R^2)$
Herzberg*	$1 - ((n-1) / (n-k-1))\ ((n-2)/(n-k-2))\ ((n+1)/n)\ (1 - R^2)$
Lord	$1 - (1 - R^2) * ((n+k+1) / (n-k-1))$

Note. All formulas presented are for fixed-effect design models except where indicated by * (random-effect design model formulas).

[a]SOS= Sum of squares

[b]v = Number of levels in a factor

[c]MS = Mean square

[d]n = Number of subjects

[e]k = Number of predictor variables

Table 3

Statistical Significance at Various Sample Sizes for a Fixed Effect Size (60%)

Source	SOS	r^2	df	MS	Fcalc	Fcrit	Decision
			n = 10				
Explained	**180.00**	**0.60**	**2**	**90.00**	5.25	4.74	REJ
Unexplain	**120.00**		7	17.143			
Total	**300.00**		9	33.333			
			n = 30				
Explained	**180.00**	**0.60**	**2**	**90.00**	20.27	3.35	REJ
Unexplain	**120.00**		27	4.44			
Total	**300.00**		29	10.345			
			n = 50				
Explained	**180.00**	**0.60**	**2**	**90.00**	35.29	3.20	REJ
Unexplain	**120.00**		47	2.55			
Total	**300.00**		49	6.122			
			n = 100				
Explained	**180.00**	**0.60**	**2**	**90.00**	72.58	3.09	REJ
Unexplain	**120.00**		97	1.24			
Total	**300.00**		99	3.03			

Note. Entries in **bold** remain fixed for the purposes of these analyses.

Table 4

Statistical Significance at Various Sample Sizes for a Fixed Effect Size (30%)

Source	SOS	r^2	df	MS	Fcalc	Fcrit	Decision
			n = 10				
Explained	**15.00**	**0.30**	**2**	**7.50**	1.50	4.74	NR
Unexplain	**35.00**		7	5.00			
Total	**50.00**		9	5.55			
			n = 30				
Explained	**15.00**	**0.30**	**2**	**7.50**	5.79	3.35	REJ
Unexplain	**35.00**		27	1.296			
Total	**50.00**		29	1.724			
			n = 50				
Explained	**15.00**	**0.30**	**2**	**7.50**	10.07	3.20	REJ
Unexplain	**35.00**		47	0.745			
Total	**50.00**		49	1.02			
			n = 100				
Explained	**15.00**	**0.30**	**2**	**7.50**	20.78	3.09	REJ
Unexplain	**35.00**		97	0.361			
Total	**50.00**		99	0.505			

Note. Entries in **bold** remain fixed for the purposes of these analyses.

Table 5

Statistical Significance at Various Sample Sizes for a Fixed Effect Size (10%)

Source	SOS	r^2	df	MS	Fcalc	Fcrit	Decision
			n = 10				
Explained	**50.00**	**0.10**	**2**	**25.00**	0.39	4.74	NR
Unexplain	**450.00**		7	64.286			
Total	**500.00**		9	55.55			
			n = 30				
Explained	**50.00**	**0.10**	**2**	**25.00**	1.50	3.35	NR
Unexplain	**450.00**		27	16.66			
Total	**500.00**		29	17.24			
			n = 50				
Explained	**50.00**	**0.10**	**2**	**25.00**	2.61	3.20	NR
Unexplain	**450.00**		47	9.57			
Total	**500.00**		49	10.20			
			n = 100				
Explained	**50.00**	**0.10**	**2**	**25.00**	5.388	3.09	REJ
Unexplain	**450.00**		97	4.64			
Total	**500.00**		99	5.05			

Note. Entries in **bold** remain fixed for the purposes of these analyses.

Table 6

Measures of Association Strength Based on Table 3 Data Using Fixed-Effect Design Model Formulas

	Measures of Association Values					
Sample Size	Eta	Omega	Epsilon	Wherry	Herzberg	Lord
10	.60	.4595	.4857	.4857	.3314	.2572
30	.60	.5621	.5704	.5704	.5274	.5120
50	.60	.5781	.5830	.5830	.5580	.5489
100	.60	.5893	.5918	.5918	.5795	.5753

Table 7

Measures of Association Strength Based on Table 4 Data Using Fixed-Effect Design Model Formulas

	Measures of Association Values					
Sample Size	Eta	Omega	Epsilon	Wherry	Herzberg	Lord
10	.30	.0910	.1000	.1000	$0_{(-.17)}$	$0_{(-.29)}$
30	.30	.2419	.2482	.2482	.1730	.1445
50	.30	.2662	.2702	.2702	.2265	.2106
100	.30	.2835	.2856	.2856	.2642	.2567

Table 8

Measures of Association Strength Based on Table 5 Data Using Fixed-Effect Design Model Formulas

	Measures of Association Values					
Sample Size	Eta	Omega	Epsilon	Wherry	Herzberg	Lord
10	.10	$0_{(-.14)}$	$0_{(-.16)}$	$0_{(-.16)}$	$0_{(-.50)}$	$0_{(-.67)}$
30	.10	.0323	.0334	.0334	$0_{(-.06)}$	$0_{(-.09)}$
50	.10	.0606	.0617	.0617	.0055	$0_{(-.02)}$
100	.10	.0807	.0814	.0814	.0540	.0443